ホームロボットにおけるクラウド型音声対話システムの利活用

○杉浦孔明，堀智織，是津耕司（情報通信研究機構）

1. はじめに

スマートフォンを始めとする種々のデバイスに音声インタフェースが導入され，広く一般に認知されるようになってきた [1, 2]．音声対話システムの分野では，開発者が容易に利用できるツールキットも公開されている（例えば [3]）．一方，人とロボットのインタラクションでは，高性能な音声認識・合成を容易に利用できる状況ではない．ロボットとの高度な音声インタラクションを可能とするためには，音声処理とロボティクスの深い知識を要求されるのが現状である．

そこで本研究では，クラウドロボティクス基盤 “rospeex”[1] を構築・公開し，サービスを長期間にわたり実運用した．音声認識および音声合成機能をクラウド化することで，音響モデルや言語モデルなどの大規模な資源をロボット上に搭載する必要がなくなり，ハードウェアを簡略化することでコストを低減できる．クラウド型音声認識・合成サービスを通じて，NICT で開発されたエンジン [2] をユーザは利用可能である．また，他のクラウド型音声認識・合成サービスに切り替えて利用することも可能である．

クラウドロボティクスやクラウドネットワークロボティクスなどの分野では，物体認識，知識共有，機械学習などのためにクラウドコンピューティングを用いるアプローチが提案されている（例えば [4]）．本研究はこれらと関連するが，ロボットの音声コミュニケーションに主眼を置く点が異なる．また，HARK [5] などミドルウェアに対応した音声コミュニケーションツールでは，内部的にスタンドアロン型のエンジンを用いている．これらのエンジンは機能的には複数言語の音声認識・合成が可能であるが，言語モデルの入れ替えなどをロボット開発者自身が行う必要がある．一方，提案手法では，次節で説明するように言語やボイスフォントの変更を簡単に行うことができる．

2. クラウドロボティクス基盤 rospeex

本節では，rospeex の機能のうち，ユーザインタフェースおよび音声認識・合成について述べる．機能の詳細については，[6] を参照されたい．図 1 左図および右図に、想定する標準的な構成および rospeex による音声対話の例を示す．

2·1 スマートフォンユーザインタフェース

図 2 に rospeex のユーザインタフェースを示す．図中の青で示す部分は発話区間として検出された部分である．このように話者に波形をリアルタイムにフィードバックすることで，声が小さいなどの問題を話者自身が気づくことができる．以下，本論文では，rospeex の使用者を「ユーザ」，ロボットとの対話者を「話者」と略す．一方，波形表示が提示されない場合，ロボットの反応から誤りの原因（発話区間検出の失敗か音声認識の誤認識か，など）を推定することは難しい．図のインタフェースでは，必要ない場合に発話区間が検出されないよう，発話区間検出機能を無効にすることも可能である．

rospeex のユーザインタフェースはブラウザ上で提供されており，マルチプラットフォームに対応している．そのため，PC のオンボードマイクや USB 接続の指向性マイクロフォン，さらにスマートフォンを入力デバイスとして用いることができる．対応ブラウザは，Google Chrome（Windows, Linux）および Firefox（Windows, Linux, Android）である．ただし、Android を OS とするハードウェアは無数に存在するため、全ての機種で動作を保証するものではない。

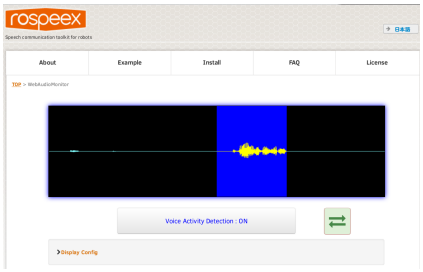

図 2 rospeex のユーザインタフェース（PC、スマートフォン）

2·2 クラウド型音声認識・合成

rospeex は複数のクラウド型音声サービスに接続可能であり，それらを切り替えて使用できる．本節では，NICT が提供する音声認識・合成サービスについて説明する．これらは，ROS を経由せずに単体としても利用可能であり，4 か国語（日英中韓）の音声認識・合成に対応している．現時点では，学術研究目的に限り無償・登録不要で公開している．本サービスでは，JSON ファイルをインタフェースとする．ユーザが用いるプログラミング言語には依存しないため，C++や Python など各種のプログラミング言語を利用可能である．

日本語の音声合成については，非モノローグ HMM 音声合成 [7] に対応している．一般的な音声合成器は人-ロボット対話に最適化されている訳ではないが，非モノローグ音声合成を選択することでロボットとの対話に特化して開発されたボイスフォントを利用可能である．

3. 実証実験

3·1 実験設定

本節では，実利用におけるロボットとの音声対話について，サーバ上のログの解析を行う．2014/1/1 から 2014/11/28 までのアクセス記録をもとに，実際の利用に

[1] rospeex は http://rospeex.org からダウンロード可能である．

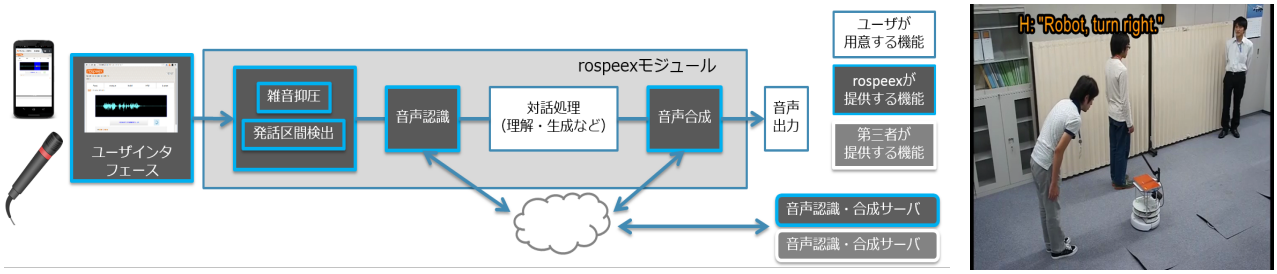

図 1 左: rospeex の構成．右: rospeex を用いた対話の例.

おける音声認識ログを解析した．本論文では rospeex の適用対象をホームロボットであると仮定し，ホームロボットに関連したカテゴリに発話を分類する．実験に用いたサーバの CPU は Intel 製 X5690（12 コア，3.47GHz），メモリは 200GB であった．

3·2 結果および考察

ログに含まれる発話の音声認識結果の総数は，44960 であった．ただし，無音など明らかに発話が含まれないものは取り除いた．このうち，頻度が 3 以上のものを分析対象として発話を分類した結果を表 1 に示す．ただし，本実験ではカテゴリを以下のように定義する．

1. 挨拶・雑談: 日常会話
   例：こんにちは，君は誰，バイバイ
2. 1 問 1 答型発話: 対話履歴を必要としない情報源への問い合わせ
   例：今何時，今日の予定を教えて，天気を教えて
3. 行動指示（移動・把持）: 移動や把持に関連する動作指示発話
   止まれ，右に折れて，本棚まで行って
4. 行動指示（家電操作）: 音声リモコンのように家電を操作する発話
   例：テレビを消して，電気をつけて
5. 行動指示（認識・学習）: センサ入力の学習または認識を指示する発話
   例：ここはどこ，あれ見て
6. 行動指示（その他）: 3-5 以外でロボットの行動を指示する発話
   例：手を上げろ，終わり
7. その他（検索・回答，判別不能）: 1-6 以外の発話．主に，質問への応答，音声認識誤りまたは判別不能な発話を含む．
   例：富士山，ちょっと

表 1 より，挨拶や雑談に比べ行動指示発話が少ないことがわかる．これは，主なユーザが開発者であり，ロボットに未実装の機能を指示しないためであると考えられる．また，約半数の発話は挨拶・雑談や 1 問 1 答型の質問である．これらの発話に対しては，一般的に提供されている質問応答や雑談対話のクラウド型 API を用いることが有効であると考えられる．

表 1 発話の分類

カテゴリ	発話数	割合 [%]
挨拶・雑談	1894	31.70
1 問 1 答型質問	1153	19.30
行動指示（移動・把持）	258	4.31
行動指示（家電操作）	229	3.83
行動指示（認識・学習）	215	3.59
行動指示（その他）	41	0.68
その他（検索・回答，判別不能）	2205	36.91
合計	5973	100

一方，行動指示はロボットごとに機能を実装する必要があり，一般的に解決は簡単ではない．音声認識誤りのため,「その他」カテゴリに分類された発話も多いと考えられるため，音声認識精度の向上は今後の課題である．さらに，対話履歴の解析を行うとともに，行動にグラウンドした対話管理が必要になるであろう．

4. おわりに

GitHub や ROS などのサービスおよびミドルウェアの普及により，ロボティクス分野においても開発したソフトウェアを公開する動きが広まっている．公開したソフトウェアの評価には，ダウンロード数，ユーザによる評価，プルリクエスト数などが用いられることが多いが，実際のインパクトを評価することは難しい．一方，本研究のようにクラウドロボティクス基盤を構築・運用することで，サーバ上の日々のアクティブユーザ数を観測することが可能になる．

本論文では，音声対話向けのクラウドロボティクス基盤 rospeex の長期実証実験とログの解析結果について述べた．本研究に関する動画は `http://rospeex.org` を参照されたい．

謝辞

本研究の一部は，科研費（若手 (B)24700188）の助成を受けて実施されたものである．

参考文献

[1] 河原達也，“音声対話システムの進化と淘汰–歴史と最近の技術動向–,” 人工知能学会誌，vol.28，no.1，pp.45–51，2013.
[2] 松田繁樹，林輝昭，葦苅豊，志賀芳則，柏岡秀紀，安田圭志，大熊英男，内山将夫，隅田英一郎，河井恒，中村哲，“多言語音声翻訳システム“ VoiceTra ”の構築と実運用による大規模実証実験,” 電子情報通信学会論文誌，vol.J96-D，no.10，pp.2549–2561，2013.
[3] 大浦圭一郎，山本大介，内匠逸，李晃伸，徳田恵一，“キャンパスの公共空間におけるユーザ参加型双方向音声案内デジタルサイネージシステム,” 人工知能学会誌，vol.28，no.1，pp.60–67，2013.
[4] K. Kamei, S. Nishio, N. Hagita, and M. Sato, “Cloud Networked Robotics,” Network, IEEE, vol.26, no.3, pp.28–34, 2012.
[5] K. Nakadai, T. Takahashi, H.G. Okuno, H. Nakajima, Y. Hasegawa, and H. Tsujino, “Design and Implementation of Robot Audition System’HARK’—Open Source Software for Listening to Three Simultaneous Speakers,” Advanced Robotics, vol.24, no.5-6, pp.739–761，2010.
[6] 杉浦孔明，堀　智織，是津耕司，“音声対話向けクラウドロボティクス基盤 rospeex の構築と長期実証実験,” 第 32 回ロボット学会学術講演会資料，pp.RSJ2014AC3I2–05，2014.
[7] K. Sugiura, Y. Shiga, H. Kawai, T. Misu, and C. Hori, “Non-Monologue HMM-Based Speech Synthesis for Service Robots: A Cloud Robotics Approach,” Proc. ICRA, pp.2237–2242, 2014.